Artificial Brains and Hybrid Minds

[Forthcoming in Müller, V. C. (ed.), *Philosophy and Theory of Artificial Intelligence 2017,* Springer.]

Artificial Brains and Hybrid Minds

Paul Schweizer

Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, Edinburgh, UK paul@inf.ed.ac.uk

Abstract The paper develops two related thought experiments exploring variations on an 'animat' theme. Animats are hybrid devices with both artificial and biological components. Traditionally, 'components' have been construed in concrete terms, as physical parts or constituent material structures. Many fascinating issues arise within this context of hybrid *physical* organization. However, within the context of functional/computational theories of mentality, demarcations based purely on *material* structure are unduly narrow. It is *abstract* functional structure which does the key work in characterizing the respective 'components' of thinking systems, while the 'stuff' of material implementation is of secondary importance. Thus the paper extends the received animat paradigm, and investigates some intriguing consequences of expanding the conception of bio-machine hybrids to include abstract functional and semantic structure. In particular, the thought experiments consider cases of mind-machine merger where there is *no* physical Brain-Machine Interface: indeed, the material human body and brain have been removed from the picture altogether. The first experiment illustrates some intrinsic theoretical difficulties in attempting to replicate the human mind in an alternative material medium, while the second reveals some deep conceptual problems in attempting to create a form of truly *Artificial* General Intelligence.

1. Introduction

In this paper I would like to explore some intriguing conceptual terrain concerning implications for future forms of mentality that might arise through advances in AI theory and technology. The discussion will proceed by examining two related variations on an 'animat' theme. Animat devices are defined as robotic machines with both active biological and artificial components (Franklin, 1995). At first pass, 'components' are most graphically construed in concrete terms, as brute material parts or constituent physical mechanisms. And of course, many intriguing issues arise within this context of hybrid physical organization, wherein biological matter such as living neural cells is coupled with engineered robotic components such as sensors and actuators. The topic is especially compelling when the biological matter in question is a living human body/brain that is augmented with technological implants and extensions. This 'cyborg' version of the bio-machine hybrid theme raises profound questions about the nature, boundaries and future of the human mind that have already stimulated much discussion (e.g. (Clark, 2003)).

However, within the overall context of Functional/Computational Theories of Mind (FCTM) (e.g. (Putnam, 1967), (Fodor, 1975) (Johnson-Laird, 1988)) central to Cognitive Science and AI, demarcations based purely on *material* structure are unduly narrow. According to FCTM, it is *abstract* functional and computational structure which does the key theoretical work in characterizing and individuating the respective 'components' of thinking agents, whereas, in accord with the principle of Multiple Realizability (MR), the 'stuff' of material implementation is generally deemed to be of secondary importance. Hence the paper aims to extend the received animat paradigm, and investigate some consequences of expanding the conception of animats and bio-machine hybrids to encompass abstract functional and semantic structure, and not just concrete physical mechanisms.

In the standard cyborg case, the core physical system is still human/biological, and this is then augmented by fusion with artificial hardware devices *via* implants, neuroprosthetics, etc. The issue then becomes one of teasing out the implications for human mentality

and identity that result from this *corporeal* blend of organic and engineered components. In these standard cases of mind/machine merger, the biological brain is *physically* impacted by other material structures, *via* Brain-Machine Interfaces, to produce a system that is no longer strictly human, but rather is a hybrid incorporating both natural and synthetic aspects.

By contrast, in the two thought experiments developed below, the *entire* physical system is synthetic – the 'brain' in question is completely artificial, and the mechanism under investigation is itself an advanced technological artifact, a robot. And instead of blending organic and synthetic *physical* parts, as in the cyborg paradigm, the scenarios below trade on the 'interaction' and comingling of purely artificial hardware systems with the *abstract* formal and linguistic structure central to organically engendered human mentality. So the hybridization involves a wholly synthetic physical device, in combination with the abstract, biologically induced cognitive and linguistic architecture of the human species. The issue then becomes one of exploring the implications for robot/human mentality that result from this hybridization – how much of the ensuing cognitive system should now be viewed as properly artificial and how much is still human?

In terms of what's essential to human mental identity, there are at least three key factors to consider. One (**1**) is internal processing structure, central to FCTM. This abstract template or computational blueprint is a defining characteristic of the human cognitive type, and it's the result of many eons of organic evolution and natural selection. Another critical feature (**2**) is our conscious experience or phenomenology: the field of occurrent, qualitative presentations (or P-consciousness, in Block's (1995) terminology). Our introspective self-identity is largely determined by this ongoing 'stream of consciousness'. And a third (**3**) key feature is the *content* of propositional attitude states such as beliefs and desires. To a great extent, who we are is dependent upon what we believe and what we want. According to the standard belief-desire framework of psychological explanation, the content of these propositional attitudes is central to

our status as rational agents, and similarly this feature is vital to Dennett's (1981) Intentional Stance.

2. An Artificial Brain as Alternate Realization

The first 'abstract animat' thought experiment begins by utilizing feature (**1**) above. According to FCTM and the attendant principle of MR, internal processing structure is not something that is essentially about our flesh and blood *embodiment*, but rather concerns a higher level description of our neurological machinery. FCTM gives rise to the mind/program analogy, wherein the mind is theoretically captured at the *software* level. So let us suppose that sometime in the sanguine future, cognitive scientists eventually discern the underlying functional/computational architecture of the human mind, and merely for the sake of convenience, let us suppose that it's some more sophisticated and far reaching version of Fodor's (1975) Language of Thought (LOT), say LOT_{37}* (clearly, this structure would have to be capable of far more than simply the manipulation of linguistically encoded propositional attitude states). So let us suppose that LOT_{37}* is the formal processing architecture which has been organically evolved and is implemented in the brain. It is the level of description which characterizes us as advanced cognitive agents, as human *minds*, and the brain is running LOT_{37}* as an indigenous formal system of rule governed symbol manipulation, in accord with the classical mind/program model.

The organic brain is then the original physical realizer of LOT_{37}*, but according to MR, our biological 'wetware' is not in principle privileged in this regard. Just as with computational procedures in general, it should be possible to take the abstract LOT_{37}* software and run it on an artificial hardware device physically quite unlike the human brain. So in the first animat scenario, let us assume that this impressive theoretical and technological feat has been accomplished – human scientists have fabricated a purely artificial electro-mechanical 'brain' that implements the human Language of

Thought. For ease of comparison, we will assume that the artificial brain occupies the cranial cavity of a fully operational robot, and hence manipulates environmental inputs and produces outputs controlling various forms of behavior in a manner completely analogous to a normal human being. Indeed, the artifact is so well crafted that it excels at some suitable version of the combined linguistic and robotic Total Turing Test (Harnad, 1991) and its success is due to the fact that the robot is an alternate realization of our own cognitive software.

Turing's original test is designed as an 'imitation game', where the goal is to *fool* someone into thinking that the computer is human. The strategy is intended to screen off anticipated (and perhaps outdated) human prejudice towards artifacts, by appealing to a standard whereby they are deemed behaviorally indistinguishable from us. But this induces a number of red-herrings, since the goal strays from detecting general intelligence *per se* to slavishly impersonating humans, warts and all (see French, 2000). In the current discussion I will therefore shift the emphasis from indistinguishability to the more salient goal of producing externally observable capabilities on a par with humans in terms of exhibiting broad-spectrum intelligence. In particular, in order to pass the presently contemplated 'soft' version of the Total Turing Test, the robot is not required to *mislead* the judges into mistaking it for a corporeal human, since this would add a myriad of restrictions and complications which are not relevant to the overall project of AI. So we will allow the judges to be cognizant of the fact that the robot is *not* physically a human. We will assume that they are suitably fair minded and impartial, and the task of the robot is to exhibit behavior that would count as appropriately intelligent in the general human case.

The robot is entirely artificial in terms of its *physical* organization and composition, but it is nonetheless a genuine case of biomechanical hybridization, since its *cognitive* architecture is an instance of the human LOT_{37}*. Thus its mechanical body and synthetic ‘central nervous system’ are advanced technological artifacts, while the abstract cognitive processing essential to its identity as a thinking agent is an organically engendered cognitive template. An artificial brain is running the software of the human mind, and in contrast to a standard cyborg case, there is now *no* biological or organic matter

present, but only the *abstract* computational structure of human cognition, which structure possesses a clearly biological as opposed to artificial etiology. According to FCTM, it is this LOT_{37}* structure which distinguishes us at the cognitive level, and we have replicated this defining human mental characteristic in an artificial brain. Hence at the salient level of description the biological brain and the robot's artificial analogue are *functionally identical*, and thus it may appear that, although the robot's synthetic brain is physically quite distinct from our own organic hardware, it nonetheless supports a purely human *mind*. However, I will now invoke feature (**2**) above to argue that this is not the whole story.

Conscious experience is a notoriously problematic topic, about which there is well known and abundant disagreement. Many theorists (e.g. Fodor) invoke FCTM only to explain the high level cognitive processing involved in propositional attitude states and rational action, while at the same time bracketing the entire issue of qualia and phenomenology. However, other authors, including (Lycan, 1987), (Jackendoff, 1987), (Johnson-Laird, 1988), (Chalmers, 1996) try to extend the reach of the computational paradigm, and contend that *conscious states* themselves arise via the implementation of the appropriate functional or information processing structure. Let us denote this extension of the basic FCTM framework 'FCTM+'.

In contrast, a primary alternative to FCTM+ contends that it is the physical substrate, the actual material *realizer* of the abstract functional structure which must be invoked in the explanation of conscious presentations (as in (Churchland, 1984)). This opposing view is a form of physicalist type-type identity theory, wherein particular material structures or processes are identified as constituting, 'giving rise to', or providing the supervenience base for the corresponding phenomenal state or property. In the case of *human* consciousness, salient aspects of the *biological brain* are thus hypothesized as responsible for various features of subjective experience, and this guides the empirical search for 'neural correlates' of consciousness and attendant psycho-physical mappings.

In comparison, a distinguishing feature of FCTM+ is that it advocates a form of 'non-reductionist' token-token physicalism motivated by the principle of Multiple Realizability. As noted above, abstract computational procedures can be implemented *via* any number

of quite distinct types of physical configuration. For example, classical Turing machines, conceived as finite programs of instructions for manipulating 0's and 1's, have the ontological status of mathematical abstractions. Like differential equations, sets, Euclid's perfectly straight lines, etc., Turing machines don't exist in *real* time or space, and they have no causal powers. In order to perform *actual* computations, an abstract Turing machine must be realized or instantiated by some suitable arrangement of mass/energy. And as Turing (1950) observed long ago, there is no privileged or unique way to do this.

The very same abstract Turing machine can be implemented *via* modern electronic circuitry, a Victorian contraption made of gears and levers (*a la* Babbage's Analytical Engine), a human being following the instructions by hand using notepad and pencil (as in the banks of clerks working at Bletchley Park), as well as more 'deviant' physical arrangements such as roles of toilet serving as the machine tape and empty beer cans for the cipher '1'. Thus there is no uniform reduction from *type* of computational state to *type* of physical state. But each particular instance or physically realized *token* of a given abstract state is still just a particular physical state or process, governed by the ordinary laws of nature. Hence the ontological commitments are held to be physicalist but non-reductivist.

The position I will now advocate is that FCTM+ is mistaken, and that qualia must supervene upon the physical substrate rather than the functional organization. Why? – because unlike computational formalisms (as well as propositional attitude states, viewed dispositionally as high level, counterfactual-supporting configurations of a computational system), conscious states are inherently *non-abstract*; they are *actual*, occurrent phenomena extended in physical time. Many qualitative presentations, such as a visual sensation of seeing a bright red dot on a display monitor in some laboratory set-up, have a measurable duration, which means that the conscious event takes place over some objectively specifiable length of time. In sharp contrast, abstract Turing machines are not extended in physical time – the computational 'steps' are not tethered to any units of physical duration and a concrete temporal dimension is entirely lacking. It is only the steps in a *materially realized* Turing machine computation that are extended in physical time, and the very same steps in different types of realization can have vastly different

temporal durations – the Analytical Engine will be markedly slower than contemporary electronic realizations.

But FCTM+ is committed to the result that qualitatively identical conscious states are maintained across wildly different kinds of physical realization, from human neural wet ware to the robot's silicon circuitry, to the gears and levers of the Analytical Engine. And this is tantamount to the claim that an actual, substantive and *invariant* qualitative phenomenon is preserved over radically diverse material systems, while at the same time, no internal physical regularities need to be preserved. But then there is no actual, occurrent factor which could serve as the causal substrate or supervenience base for the substantive and invariant phenomenon of internal conscious experience. The advocate of FCTM+ cannot plausibly rejoin that it is invariance of *formal* or *functional* role which supplies this basis, since formal role is abstract, and such abstract features can only be implemented *via* actual properties, but they do not have the power to *produce* them (see Schweizer, 2002) for related discussion).[1]

Indeed, physical conservation laws hold that all physical events must have a purely physical cause. So if one is really a physicalist (as opposed to some sort of crypto-dualist) and holds that occurrent qualitative experience is an actual event rather than a mere abstraction, then it follows that the cause must be physical. Hence it would seem to be entailed by basic conservation laws that the material brain (natural or synthetic) must do the causal work of the mind. If internal conscious states are real phenomena extended in time, then their ultimate source must be the brain/hardware – they must depend upon intrinsic properties of the *realizer* as a proper subsystem of the actual world.

Conscious experiences are then seen as hardware states that play an abstract functional role. This abstract role remains a legitimate software concern, and it must be preserved across divergent realizations. But the actual properties of consciousness are a feature of the material substrate, and (unless one has some sort of 'magical' theory of computation, whereby implementing a computational formalism

[1] Thus the critique applies not just to classical computation and the mind/program model, but to any approach committed to abstract structural explanation and MR, such as connectionist architectures.

somehow imbues a physical system with mysterious powers and properties over and above its ordinary physical traits) these are not guaranteed to be preserved across widely different physical systems. Qualitative aspect is essentially conditioned by the hardware and hence *is* largely a matter of our flesh and blood embodiment (the above is comparable to some of the views put forward by Searle (1992), although here I am making a claim about qualitative states *simplicitor*, and no assertion whatever about 'Intentionality').

My position is not in direct conflict with the functionalist-driven view that some advanced functional roles such as a self-model and other meta-cognitive features may *require* conscious implementation, and hence that alternative realizations of the human cognitive template that are purportedly devoid of conscious experience (e.g. Block's (1978) 'Chinese Nation') are not genuine possibilities in the first place, and hence cannot serve as hypothetical counterexamples to FCTM+. Although I am in principle agnostic as to whether *any* functional role is such that it requires conscious presentations, even if this constraint on possible implementations is granted, it does not follow that the supporting phenomenology must be *qualitatively identical* to ours. There is no reason to suppose that the field of human conscious experience is the *unique solution* to the functional constraints imposed by LOT_{37}*, and hence the purely abstract structures of FCTM+ are not sufficient to determine our particular phenomenology.

Qualitative presentations in the case of, e.g., visual perception, play the functional role of providing information about the external environment. Hence LOT_{37}* would include functional pathways for processing sensory inputs and utilizing this information for executive control, such as locomotion and navigation. However this is not enough to determine the qualitative aspects produced by the physical vehicle that implements this role. So the very same abstract LOT_{37}* functional specification, when implemented in human neurophysiology, will result in qualitatively different phenomenology than when implemented in the robot's silicon circuitry. The physical vehicle implementing the functional role will determine the actual and occurrent aspects of qualitative experience, and indeed, physiological variations between human individuals could well induce large disparities in qualia, even among con-specifics.

It's not clear how deeply the LOT_{37}* high-level rational and linguistic processing, integrated with more ancient perceptual and navigational architecture, will penetrate into the qualitatively manifested differences in the robot's physical substrate. Thus if employ some version of Dennett's (1992, 2003) 'heterophenomenology', it's conceivable that the robot could report on qualitative aspects of its conscious states that diverge from ours, and hence would cause it to fail a Total Turing Test designed as a strict 'imitation game'. But in the context of the present thought experiment we will assume that the robot is indeed functionally isomorphic, and hence its verbal outputs reveal no qualitative differences, as in the case of qualitative variations between con-specifics not revealed through verbal outputs.

There are a number of different types of conscious states, including perceptual, cognitive, and affective. The LOT_{37}* will encompass high-level rational and linguistic processing, integrated with more ancient perceptual and navigational architecture, and these will be realized by divergent physical media in the robotic 'cognitive clone'. However, affective conscious states, such as moods and emotions, are far less obviously tethered to *abstract* processing structure, and much more directly related to brute biochemical influences. Since the robot is a synthetic device, it will not possess human hormones, and thus we should expect its affective states (if any) to be qualitatively distinct from ours.

And as a final consideration in this vein, it's relevant to mention 'noise' as yet a fourth type of conscious experience. For example, when I rub my knuckles against my closed eyes and then open them, I see various twinkling yellow spots. As with the tingling sensation of a sneeze, etc., I would take these phenomena to serve no functional role whatever, but rather to be mere noise in my organic hardware system – just evolutionary spandrels. And since the robot's hardware is fundamentally different, it seems reasonable to conclude that it would not experience qualitatively identical forms of noise.[2]

In any case, it's quite safe to say that consciousness is still deeply mysterious, and currently no one has a firmly established or conceptually complete and satisfactory account. Given our present state of

[2] There is nothing to prevent an FCTM+ advocate from attributing a functional role to tingles, afterimages, etc., but I would view this as merely an ad hoc strategy for defending their theory against an obvious objection.

insipient understanding, it's inevitable that conjecture and speculation will abound. Neither the FCTM+ view nor the contrasting hardware based account defended above are yet confirmed (or definitively refuted). I've offered some criticisms of the functionalist view and various reasons for favoring the hardware based approach, but it nonetheless remains an open question. And the focus of the current exercise is a thought experiment, rather than an attempt to conclusively establish the truth of some proposition. So for the purposes of the thought experiment, those who might still adhere to FCTM+ are invited to construe the results in a conditional format – *if* the (speculative) hardware based account turns out to be correct and the (speculative) FCTM+ view turns out to be false, *then* the consequent follows. Hence for the sake of argument we will now proceed on the assumption that the antecedent of the foregoing conditional turns out to be true.

So although the robot's mind is *functionally* identical to that of a human, we should expect its *synthetic phenomenology* to be highly divergent, since the substrate in which the functional structure is realized is very different than human neurophysiology. The hypothetical robot brain sustains a form of artificial consciousness that is qualitatively distinct from ours, and potentially very alien. Thus to the extent that phenomenology is a constituent of general mentality, the robot mind is distinct from the human mind. The LOT_{37}* processing structure is the result of many eons of organic evolution and natural selection, and in this respect the robot's cognitive architecture has a clearly biological etiology. But even though the robot's abstract mental processing structure is quintessentially human, its conscious experience is artificial, and is qualitatively dissimilar to ours. Hence the overall type of mind induced is not purely human, but rather is a bio-machine *hybrid*.

3. An Artificial Brain Implementing Synthetic Cognitive Architecture

The second scenario takes yet a further step of abstraction. In the first case we detached computational structure from underlying hardware, and exploited MR to yield an artificial realization of the biologically evolved LOT_{37}*. Now we will abstract away from *internal* factors altogether, including *both* physical substrate and the cognitive software it's running. We consider a computational artifact again capable of passing the combined linguistic and robotic Total Turing Test, but where the robot's internal processing structure is now entirely artificial. The robot's cognitive architecture has been custom designed by AI researchers, and is *functionally* as unlike LOT_{37}* as the first robot's artificial 'brain' is *physically* unlike human neurophysiology. This is a case of successfully manifesting all aspects of intelligent human behavior in the shared linguistic and spatiotemporal environment, but where this is achieved *via* an internal processing structure vastly different from humans.

In response to the question above – how much of the robot's mind should be viewed as properly artificial? – it may appear that in this case the answer should be '*all of it*', that we have succeeded in producing a truly synthetic mind, comprised of both artificial software and an artificial brain (and perhaps replete with synthetic phenomenology, as in the previous thought experiment). However, I will now invoke feature (**3**) to argue that, again, this is not the whole story.

The *content* of propositional attitude states such as beliefs and desires is surely a core feature of minds. As above, to a great extent, our mental identity is dependent upon what we believe and what we want. According to the standard belief-desire framework of psychological explanation, the content of these propositional attitudes is central to our status as *rational agents* and similarly this feature is vital to Dennett's Intentional Stance. And this is important, because the issue at hand does not concern the bare mechanical and engineering factors involved in designing and building a robot able to pass the Total Turing Test. Instead the issue concerns the subsequent *evaluation* of the artifact with respect to its semantic and intentional properties, including genuine intelligence, understanding, reference

for its assorted linguistics outputs, and the attribution of associated mental states, such as *believing* that snow is white, *knowing* that water is H_2O, *wanting* to pass the Total Turing Test, etc.

As in the case of behaviorally indistinguishable *humans,* the robot will be evaluated as an Intentional System harboring assorted beliefs, desires and other intentional states, and whose behavior can be explained and predicted on the basis of the *content* of these states. Accordingly, the robot's salient sonic emissions are *interpreted* as asserting various propositions and expressing assorted cognitive contents. For example, suppose Robbie the Robot, our hypothetical Total Turing Test artifact, is ambling down a path and there's a fallen log in the way. Robbie lifts his artificial leg unusually high and steps over the log. When asked 'Why did you lift your leg so high?' Robbie emits the rejoinder 'I saw the fallen log and did not want to trip on it.' Robbie is reporting the content of his relevant propositional attitude states in English, and if we are to interpret him as such then they depend in an essential manner on the public, externally determined semantics for this *human* Natural Language (NL). This is entailed if we are to construe the artifact as a rational agent, as the locus of some genuine form of mentality, and hence as *using* NL in a meaningful and referential manner, rather than just *mentioning* syntactic strings generated by its internal linguistic processing system.

According to Putnam's (1975) highly influential and compelling analysis, the semantics of NL 'ain't in the head' of any individual human agent, but rather are set by the encompassing sociolinguistic community of which the agent is a member. But if linguistic meanings ain't in the head of any individual humans, then they surely ain't in the data base of Robbie the Robot. As originally propounded by Burge (1979), Putnam's semantic externalism for NL implies that *mental content* is non-individualistic. The propositional attitudes of human individuals derive their meaning from the engulfing sociolinguistic medium.

And just as in the case of individual human mentality, so too for Robbie. The Total Turing Test robot is inextricably embedded in a *human* sociolinguistic community with its associated network of *human* cognitive scaffolding (Rupert, 2009). Thus because of semantic externalism and the ineleminable role of the *biologically* en-

gendered sociolinguistic medium, the robot's wide *mental content* will be derived from the embedding Natural Language culture. Even though Robbie is wholly non-natural as an isolated artifact, his *wide* propositional attitudes will be no more artificial than yours or mine, and his essential status as an Intentional System is dependent on the *human* sociolinguistic culture in which he functions (see Schweizer, 2012 for further discussion).

So to the extent that he's a rational agent susceptible to the framework of Belief-Desire explanation and prediction, Robbie's mind has an ineliminable human component. Natural languages have evolved over many cycles of adaptation and selection, and in this sense the sociolinguistic context upon which Robbie's mental content depends can be seen as an 'organic' component with a clearly biological etiology. Hence such a robot would not be a case of purely artificial mentality, but rather a complex blend of artificial internal processing structures in conjunction with biologically engendered sociolinguistic content. In the general case, mental states are determined both by *internal* factors, such as computational processing configurations and phenomenology, as well as *external* factors, such as the wide propositional content of beliefs and desires. Hence the robot is a bio-machine hybrid in terms of its external *versus* internal facets of mentality.

4. Conclusion

The discussion has extended the received animat paradigm by exploring two cases of genuine mind-machine merger, but where there is *no* physical Brain-Machine Interface – indeed, the material human body/brain has been removed from the picture altogether. The first thought experiment utilizes FCTM and the attendant principle of MR to envision a case where the quintessentially human LOT_{37}* functional/computational architecture is implemented in a humanoid artifact. The widely embraced mind/program analogy would seem to imply that the resulting 'cognitive clone' would possess a purely *human* mind, sustained by an alternative physical substrate. However,

it is argued that the situation is not so straightforward, and that the artificial consciousness induced by the robot's divergent hardware would result in a type of mentality not purely human, but rather a form of bio-machine hybrid. And this illustrates some intrinsic theoretical difficulties in attempting to replicate the human mind in an alternative material medium.

In the second thought experiment, the human body/brain as well as its organically engendered cognitive architecture have been removed, and the robot in question runs custom designed artificial software. Nonetheless, its status as an Intentional System, and the attendant content of its propositional attitude states is essentially human, which illustrates some deep theoretical difficulties in attempting to create a form of purely *Artificial* General Intelligence, a truly *artificial mind*.[3]

Acknowledgments I would like to thank the reviewers Chuanfei Chin, Sam Freed and Dagmar Monett for useful comments.

References

Block, N. (1978). Troubles with functionalism, in C. W. Savage (ed.), *Perception and Cognition*, Minneapolis: University of Minnesota Press.

Block, N. (1995). On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, **18**, pp. 227-247.

Burge, T. (1979). Individualism and the mental. In French, P., Euhling, T., and Wettstein, H. (eds.), *Studies in Epistemology*, vol. 4, *Midwest Studies in Philosophy*, University of Minnesota Press.

Chalmers, D. (1996). *The Conscious Mind*, Oxford: OUP.

Churchland, P. (1984). *Matter and Consciousness*, Cambridge: MIT Press.

Clark, A. (2003). *Natural-Born Cyborgs*, Oxford: OUP.

[3] In view of externalist semantical implications for mental content, a fully artificial form of mentality would require a 'Planet of the Robots' scenario, a community of feral artifacts not programmed with human NL. Instead, the robots would need to evolve their own sociolinguistic community from scratch, just as the human race did. In this manner the robotic mental states and contents would be *genuinely artificial*, just as advanced biological creatures on another planet would possess a *genuinely alien* form of mentality.

Dennett, D. (1981). True believers: the intentional strategy and why it works. In A. F. Heath (Ed.) *Scientific Explanation: Papers Based on Herbert Spencer Lectures given in the University of Oxford*, Oxford: University Press.

Dennett, D. (1992). *Consciousness Explained*, London: Viking Press.

Dennett, D. (2003). Who's on first? Heterophenomenlogy explained, *Journal of Consciousness Studies*, **10**, pp. 19-30.

Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.

Franklin, S. (1995). *Artificial Minds*, Cambridge: MIT Press.

French, R. (2000). The Turing test: the first 50 years, *Trends in Cognitive Sciences* 4: 115-122.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem, *Minds and Machines*, **1**, 43-54.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*, Cambridge: MIT Press.

Johnson-Laird, P.N. (1988). *The Computer and the Mind*, Cambridge: Harvard University Press.

Lycan, W. G. (1987). *Consciousness*, Cambridge: MIT Press.

Putnam, H. (1967). Psychological predicates. In W.H. Capitan and D. D. Merrill (Eds.) *Art, Mind and Religion*, Pittsburgh: University of Pittsburgh Press.

Putnam, H. (1975). The meaning of 'meaning'. In *Mind, Language and Reality*, Cambridge University Press.

Rupert, R. (2009). *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Schweizer, P. (2002). Consciousness and computation: A reply to Dunlop.' *Minds and Machines* 12, 143-144.

Schweizer, P. (2012). The externalist foundations of a truly total Turing test, *Minds and Machines*, **22**(3),191-212.

Searle, J. (1992). *The Rediscovery of the Mind*, Cambridge: MIT Press.